Projekt analityczny

Weronika Bola

12 czerwca 2018

Opis danych

Analizowane dane dotyczą opóźnień połączeń lotniczych w USA w lipcu 2017 r.

Liczba zawartych tabel w bazie danych wynosi 4. Przedstawione są w nich informacje o nazwach linii lotniczych, nazwach lotnisk, opóźnieniach lotów, pogodzie i dniach tygodnia.

Dane zawarte w atrybutach tych tabel, które są brane pod uwagę, odnoszą się do: opóźnień przylotów i wylotów w minutach, informacji o przewoźnikach, z których korzystają pasażerowie, zarówno miastach wylotów, jak i przylotów, nazw lotów oraz dat poszczególnych lotów wraz z dniami tygodnia, w których były wykonywane.

Połączenie

```
library(odbc)
con <- DBI::dbConnect(odbc(),
                      Driver   = "SQL Server",
                      Server   = "mssql-2016.labs.wmi.amu.edu.pl",
                      Database = "dbad_flights",
                      Port     = 1433)
```

Pytanie 1.

Jakie było średnie opóźnienie przylotu?

```
dbGetQuery(con, "SELECT AVG(arr_delay) [średnie_opóźnienie_przylotu]
                 FROM Flight_delays")
```

```
##   średnie_opóźnienie_przylotu
## 1                     8.31161
```

Pytanie 2.

Jakie było maksymalne opóźnienie przylotu?

```
dbGetQuery(con, "SELECT MAX(arr_delay) [maksymalne_opóźnienie_przylotu]
                 FROM Flight_delays")
```

```
##   maksymalne_opóźnienie_przylotu
## 1                           1895
```

Pytanie 3.

Który lot miał największe opóźnienie przylotu?

[przewoźnik, miasto wylotu, miasto przylotu, data lotu, opóźnienie]

```
dbGetQuery(con, "SELECT arr_delay [opóźnienie],
                       carrier [przewoźnik],
                       origin_city_name [miasto_wylotu],
                       dest_city_name [miasto_przylotu],
                       fl_date [data_lotu]
                 FROM Flight_delays
                 WHERE arr_delay = (SELECT MAX(arr_delay)
                                    FROM Flight_delays)")
```

```
##   opóźnienie przewoźnik miasto_wylotu miasto_przylotu  data_lotu
## 1       1895         AA      Kona, HI Los Angeles, CA 2017-07-26
```

Pytanie 4.

Które dni tygodnia są najgorsze do podróżowania?

[tabela zawierająca dla każdego dnia tygodnia średni czas opóźnienia]

```
dbGetQuery(con, "SELECT CAST(w.weekday_name AS VARCHAR(20)) [dzień_tygodnia],
                       AVG(f.arr_delay) [średni_czas_opóźnienia]
                 FROM Flight_delays f
                      JOIN Weekdays w
                        ON f.day_of_week=w.weekday_id
                 GROUP BY CAST(w.weekday_name AS VARCHAR(20))
                 ORDER BY [średni_czas_opóźnienia] DESC")
```

```
##   dzień_tygodnia średni_czas_opóźnienia
## 1         Friday              14.452013
## 2         Monday              10.576251
## 3       Thursday               8.507184
## 4      Wednesday               8.457371
## 5       Saturday               7.545721
## 6        Tuesday               4.619250
## 7         Sunday               4.261949
```

Pytanie 5.

Które linie lotnicze latające z San Francisco (SFO) mają najmniejsze opóźnienia przylotu?

[tabela zawierająca nazwę przewoźnika oraz średnie opóźnienie z jego wszystkich lotów]

```
dbGetQuery(con, "SELECT *
                 FROM
                 (
                    SELECT AVG(dep_delay+arr_delay) AS [średnie_opóźnienie],
                           CAST(carrier AS VARCHAR(100)) AS [przewoźnik]
                    FROM Flight_delays
                    WHERE origin LIKE 'SFO'
                    GROUP BY CAST(carrier AS VARCHAR(100))
                 ) t
                 WHERE [średnie_opóźnienie] <= ALL
                 (
                    SELECT AVG(dep_delay+arr_delay) AS [średnie_opóźnienie]
                    FROM Flight_delays
```

```
                    WHERE origin LIKE 'SFO'
                    GROUP BY CAST(carrier AS VARCHAR(100))
                )")
```

```
##   średnie_opóźnienie przewoźnik
## 1                 -5         HA
```

Pytanie 6.

Jaka część linii lotniczych ma regularne opóźnienia, tj. jej lot ma średnio co najmniej 10 min. opóźnienia?

[tylko linie lotnicze występujące w tabeli Flight_delays]

```
x <- dbGetQuery(con, "SELECT *
                      FROM
                      (
                         SELECT AVG(arr_delay) AS [opóźnienie],
                         CAST(carrier AS VARCHAR(100)) AS [linia_lotnicza]
                         FROM Flight_delays
                         GROUP BY CAST(carrier AS VARCHAR(100))
                      ) AS t
                      WHERE opóźnienie >= 10")
y <- dbGetQuery(con, "SELECT DISTINCT CAST(carrier AS varchar(100))
                      FROM Flight_delays")
(wynik <- nrow(x)/nrow(y))
```

```
## [1] 0.3333333
```

Pytanie 7.

Jak opóźnienia wylotów wpływają na opóźnienia przylotów?

[współczynnik korelacji Pearsona między czasem opóźnienia wylotów a czasem opóźnienia przylotów]

```
w <- dbGetQuery(con, "SELECT dep_delay [czas_opóźnienia_wylotów]
                      FROM Flight_delays")
p <- dbGetQuery(con, "SELECT arr_delay [czas_opóźnienia_przylotów]
                      FROM Flight_delays")
w <- w[,1]
p <- p[,1]
w[is.na(w)] <- 0
p[is.na(p)] <- 0
cor(w, p)
```

```
## [1] 0.9597573
```

Pytanie 8.

Która linia lotnicza miała największy wzrost (w wartościach bezwzględnych) średniego opóźnienia przylotów w ostatnim tygodniu miesiąca, tj. między 1-23 a 24-31 lipca?

[nazwa przewoźnika oraz wzrost]

```
a <- dbGetQuery(con, "SELECT AVG(f.arr_delay) [średnie_opóźnienie_przylotóW],
                       RIGHT(CAST(a.airline_name AS VARCHAR(100)),2)
                     FROM Flight_delays f
                          JOIN Airlines a
                            ON a.airline_id=f.airline_id
                     WHERE (MONTH(CAST(f.fl_date AS VARCHAR(100)))=7)
                            AND (DAY(CAST(f.fl_date AS VARCHAR(100))) BETWEEN 01 AND 23)
                     GROUP BY CAST(a.airline_name AS VARCHAR(100))
                     ORDER BY CAST(a.airline_name AS VARCHAR(100))")
b <- dbGetQuery(con, "SELECT AVG(f.arr_delay) [średnie_opóźnienie_przylotóW],
                       RIGHT(CAST(a.airline_name AS VARCHAR(100)),2)
                     FROM Flight_delays f
                          JOIN Airlines a
                            ON a.airline_id=f.airline_id
                     WHERE (MONTH(CAST(f.fl_date AS VARCHAR(100)))=7)
                            AND (DAY(CAST(f.fl_date AS VARCHAR(100))) BETWEEN 24 AND 31)
                     GROUP BY CAST(a.airline_name AS VARCHAR(100))
                     ORDER BY CAST(a.airline_name AS VARCHAR(100))")
c=data.frame(przewoznik=a[,2],wzrost=abs(a[,1]-b[,1]))
c[which.max(c[,2]),]
```

```
##   przewoznik   wzrost
## 4         EV 11.12226
```

Pytanie 9.

Które linie lotnicze latają zarówno na trasie SFO -› PDX (Portland), jak i SFO -› EUG (Eugene)?

```
dbGetQuery(con, "SELECT airline_name [linia_lotnicza]
                 FROM Airlines
                 WHERE airline_id IN (
                                        SELECT airline_id
                                        FROM Flight_delays
                                        WHERE (origin LIKE 'SFO' AND dest LIKE 'PDX')
                                      )
                 AND airline_id IN (
                                        SELECT airline_id
                                        FROM Flight_delays
                                        WHERE (origin LIKE 'SFO' AND dest LIKE 'EUG')
                                    )")
```

```
##              linia_lotnicza
## 1 United Air Lines Inc.: UA
## 2 SkyWest Airlines Inc.: OO
```

Pytanie 10.

Jak najszybciej dostać się z Chicago do Stanfordu, zakładając wylot po 14:00 czasu lokalnego?

[tabela zawierająca jako miejsce wylotu Midway (MDW) lub O'Hare (ORD), jako miejsce przylotu San Francisco (SFO), San Jose (SJC) lub Oakland (OAK) oraz średni czas opóźnienia przylotu dla wylotów po 14:00 czasu lokalnego (atrybut crs_dep_time); wyniki pogrupowane po miejscu wylotu i przylotu, posortowane malejąco]

```
dbGetQuery(con, "SELECT AVG(arr_delay) [średni_czas],
                       CAST(origin AS VARCHAR(8)) [miejsce_wylotu],
                       CAST(dest AS VARCHAR(8)) [miejsce_przylotu]
                 FROM Flight_delays
                 WHERE (CAST(origin AS VARCHAR(100)) IN ('MDW','ORD'))
                       AND (CAST(dest AS VARCHAR(100)) IN ('SFO', 'SJC', 'OAK'))
                       AND (LEN(CAST(crs_dep_time AS VARCHAR(5)))=4)
                       AND (CAST(LEFT(CAST(crs_dep_time AS VARCHAR(5)),2) AS INT)>=14)
                 GROUP BY CAST(origin AS VARCHAR(8)), CAST(dest AS VARCHAR(8))
                 ORDER BY średni_czas")
```

```
##   średni_czas miejsce_wylotu miejsce_przylotu
## 1    2.064516            ORD              OAK
## 2    4.758065            MDW              OAK
## 3    7.311111            ORD              SJC
## 4   10.600000            MDW              SJC
## 5   14.357143            ORD              SFO
## 6   15.114286            MDW              SFO
```

Podsumowanie

Z przedstawionych danych wynika, że średnie opóźnienie przylotu wynosiło 8.31161, a maksymalne było równe 1895.

Największe opóźnienie przylotu miał lot, w którym przewoźnikiem były amerykańskie linie lotnicze, miastem wylotu byla Kona, natomiast przylotu Los Angeles. Lot odbył się 26 lipca.

Najgorszymi dniami do podróżowania były w kolejności: piątek, poniedziałek, czwartek, środa, sobota, wtorek, niedziela.

Liniami lotniczymi latającymi z San Francisco, które miały najmniejsze opóźnienia przylotu były Hawaiian Airlines Inc.

1/3 linii lotniczych miała regularne opóźnienia, tj. ich lot miał średnio co najmniej 10 min. opóźnienia.

Opóźnienia wylotów miały duży wpływ na opóźnienia przylotów, wraz z ich wzrostem powiększały się opóźnienia przylotów.

Największy wzrost średniego opóźnienia przylotów w ostatnim tygodniu miesiąca, tj. między 1-23 a 24-31 lipca miała linia lotnicza ExpressJet Airlines Inc.

Liniami lotniczymi, które latały zarówno na trasie San Francisco -> Portland, jak i San Francisco -> Eugene były United Air Lines Inc. i SkyWest Airlines Inc.

Z Chicago do Stanfordu zakładając wylot po 14:00 czasu lokalnego najszybciej dostać się wylatując z O'Hare, a przylatując do Oakland.